ITESO, Universidad Jesuita de Guadalajara

InstitutoTecnológico y de Estudios Superiores de Occidente

Maestría Ciencia de Datos

Análisis Estadístico Multivariado

Segundo Examen Parcial

Estudiante: Daniel Nuño
Profesor: Dra. Rocío Carrasco
Fecha entrega: Noviembre 21, 2021

---

1. Regresión Lineal Simple y Múltiple para Seoul Bike

a) ¿Qué supuestos debe cumplir un modelo que describa la relación lineal entre dos variables? Describa en qué consiste cada uno de ellos.

- Linealidad: Una variable o conjunto de variables esta correlacionada y puede describir a la otra.
- Homocedasticidad: La varianza de los errores debe ser la mismo, es decir, que el ajuste es igual de preciso independientemente de los valores que tome la variable independiente
- Normalidad: Para cada valor de la variable independiente, los residuos ei tienen distribución normal de media cero.
- Independencia: Autocorrelación. Los residuos deben ser independientes entre sí.

b) Para el modelo de regresión lineal simple, ¿cuál es la variable independiente y cuál es la variable dependiente? (según la base de datos que eligieron)
Las variables dependientes es la cantidad de bicicletas rentadas, la variable independientes puede ser temperatura(c).

c) Escriba un enunciado planteando el objetivo o el problema a resolver.
Es importante que la bicicleta de alquiler esté disponible y sea accesible para el público en el momento adecuado, ya que disminuye el tiempo de espera. Con el tiempo, proporcionar a la ciudad un suministro estable de bicicletas de alquiler se convierte en una gran preocupación. La parte crucial es la predicción de las bicicletas requeridas.

d) Obtenga el modelo de regresión simple y escriba su ecuación.
Voy a hacer algunas observaciones para el procesamiento de datos.

- Cuando Functioning Day es No, no hay renta de bicicletas. Voy a quitarla de los datos.
- La hora es importante para determinar la cantidad de bicicletas, usualmente no hay bicicletas rentadas en la madrugada.
- Un *buen clima* es determinante.

debajo del análisis exploratorio esta la ecuación

In [136…

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import seaborn as sns
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
from scipy import stats
from statsmodels.compat import lzip
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from statsmodels.formula.api import ols
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import statsmodels.stats.multicomp as mc
```

In [85]:
```
df = pd.read_csv("C:/Users/nuno/OneDrive - ITESO/Ciencia de Datos/"
                 "analisis_estadistico_multivariado/code/SeoulBikeData.csv",
                 encoding = 'unicode_escape')
to_drop = df.index[df["Functioning Day"] == "No"].tolist()
df.drop(to_drop, axis=0, inplace = True)
df.drop(['Functioning Day'], axis=1, inplace=True)
df['Holiday'].replace('No Holiday', 0, inplace=True)
df['Holiday'].replace('Holiday', 1, inplace=True)
df['xhr'] = np.sin(2*np.pi*df['Hour']/24)
df['yhr'] = np.cos(2*np.pi*df['Hour']/24)
df.head()
```

Out[85]:

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Ra
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	

In [86]:
```
df.dtypes
```

Out[86]:
```
Date                          object
Rented Bike Count              int64
Hour                           int64
Temperature(°C)              float64
Humidity(%)                    int64
Wind speed (m/s)             float64
Visibility (10m)               int64
Dew point temperature(°C)    float64
Solar Radiation (MJ/m2)      float64
Rainfall(mm)                 float64
Snowfall (cm)                float64
Seasons                       object
Holiday                        int64
xhr                          float64
yhr                          float64
dtype: object
```

In [87]:
```
df.describe()
```

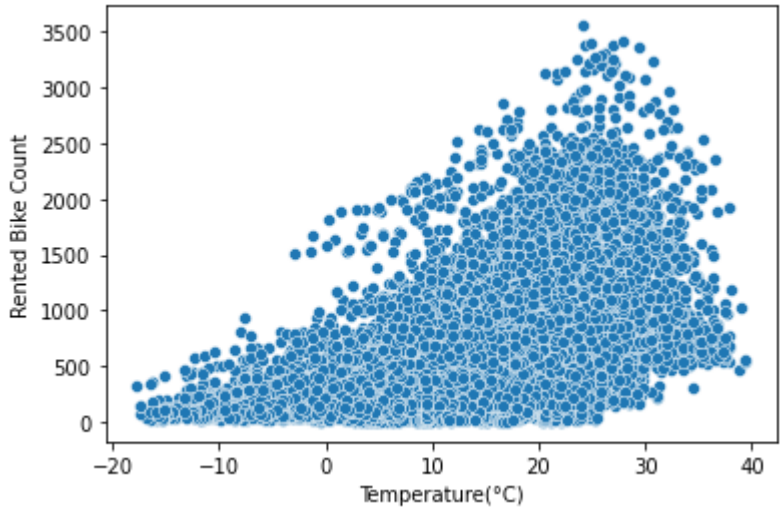
Out[87]:

	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	8
count	8465.000000	8465.000000	8465.000000	8465.000000	8465.000000	8465.000000	8465.000000	
mean	729.156999	11.507029	12.771057	58.147194	1.725883	1433.873479	3.944997	
std	642.351166	6.920899	12.104375	20.484839	1.034281	609.051229	13.242399	

	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)
min	2.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000
25%	214.000000	6.000000	3.000000	42.000000	0.900000	935.000000	-5.100000
50%	542.000000	12.000000	13.500000	57.000000	1.500000	1690.000000	4.700000
75%	1084.000000	18.000000	22.700000	74.000000	2.300000	2000.000000	15.200000
max	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000

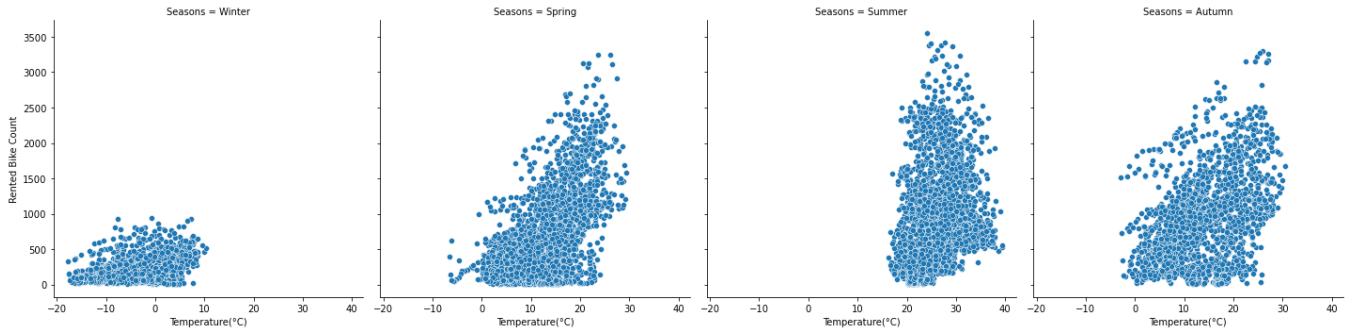
In [88]:
```
sns.scatterplot(x=df['Temperature(°C)'], y=df['Rented Bike Count'])
```

Out[88]:
```
<AxesSubplot:xlabel='Temperature(°C)', ylabel='Rented Bike Count'>
```



In [89]:
```
sns.relplot(
    data=df, x='Temperature(°C)', y='Rented Bike Count',
    col='Seasons', kind='scatter'
)
```

Out[89]:
```
<seaborn.axisgrid.FacetGrid at 0x259da8453a0>
```



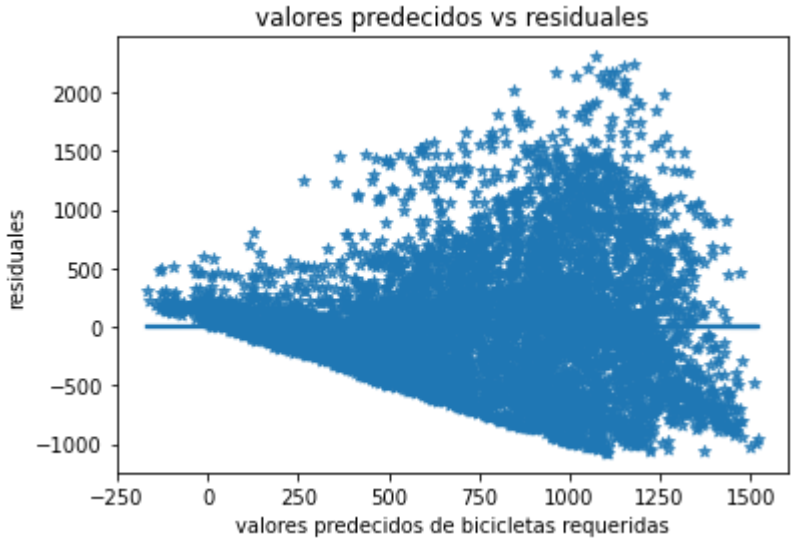
In [90]:
```
corr_mtrx = df.corr()
corr_mtrx.iloc[:,0]
```

Out[90]:
```
Rented Bike Count            1.000000
Hour                         0.425256
Temperature(°C)              0.562740
Humidity(%)                 -0.201973
Wind speed (m/s)             0.125022
Visibility (10m)             0.212323
Dew point temperature(°C)    0.400263
Solar Radiation (MJ/m2)      0.273862
Rainfall(mm)                -0.128626
Snowfall (cm)               -0.151611
Holiday                     -0.070070
xhr                         -0.447846
yhr                         -0.102584
Name: Rented Bike Count, dtype: float64
```

In [91]:
```
x = np.array(df['Temperature(°C)'])
y = np.array(df['Rented Bike Count'])
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2)
x_train= x_train.reshape(-1, 1)
y_train= y_train.reshape(-1, 1)
x_test = x_test.reshape(-1, 1)
y_test = y_test.reshape(-1, 1)
rls = linear_model.LinearRegression()
rls.fit(x_train, y_train)
y_pred = rls.predict(x_test)
y_pred2 = rls.predict(x_train)
print("b_1", rls.coef_)
print("b_0", rls.intercept_)
print("r-squared", rls.score(x_train, y_train))

x2 = sm.add_constant(x_train, prepend=True)
rls2=sm.OLS(endog=y_train, exog=x2)
rls2=rls2.fit()
print(rls2.summary(), "\n")
y_pred3 = rls2.predict(x2)
error2 = y_train.reshape(-1,1) - y_pred3.reshape(6772,1)
names=['Lagrange multiplier statistic',
        'p-value',
        'f-value',
        'f p-value']
test = sms.het_breuschpagan(error2, x2)
print(lzip(names, test), "\n")

plt.figure()
sns.regplot(x=y_pred3, y=error2, marker='*')
plt.xlabel('valores predecidos de bicicletas requeridas')
plt.ylabel('residuales')
plt.title('valores predecidos vs residuales')
plt.show()
```

```
b_1 [[29.60711718]]
b_0 [352.35625611]
r-squared 0.31581610191862597
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.316
Model:                            OLS   Adj. R-squared:                  0.316
Method:                 Least Squares   F-statistic:                     3125.
Date:                Sun, 21 Nov 2021   Prob (F-statistic):               0.00
```

```
Time:                        18:19:18   Log-Likelihood:                -52072.
No. Observations:                6772   AIC:                         1.041e+05
Df Residuals:                    6770   BIC:                         1.042e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        352.3563      9.303     37.876      0.000     334.120     370.593
x1            29.6071      0.530     55.902      0.000      28.569      30.645
==============================================================================
Omnibus:                      756.551   Durbin-Watson:                   2.027
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1137.846
Skew:                           0.827   Prob(JB):                    8.31e-248
Kurtosis:                       4.138   Cond. No.                         25.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[('Lagrange multiplier statistic', 829.4639689940991), ('p-value', 2.119370803325487e-182),
('f-value', 944.9620567364931), ('f p-value', 2.2902364880317233e-194)]
```



d) La ecuación lineal es $y = \beta_0 + \beta_1 x$ y para la base de datos considerando temperatura como la variable independiente la ecuación es $RentedBikeCount = 347.02 + (30.05)Temperature$

e) ¿Por qué es importante separar los datos en 80% para entrenamiento y 20% para prueba?
Es importante evaluar la precisión del modelo. El porcentaje de prueba son valores desconocidos para el modelo, pero conocidos por nosotros por lo que podemos calcular lo bien que lo hizo. También sirve para evitar sobreajuste.

f) De una interpretación de los resultados obtenidos (Summary)

R-squared es el coeficiente de determinación. Indica cuanto de la variable independiente es explicada por cambios en nuestra variable dependiente. En términos de porcentaje 0.31 significa que nuestro modelo explica 31% del cambio de las bicicletas rentadas. **Adj. R-squared** es importante para analizar la eficacia de múltiples variables dependientes en el modelo. La regresión lineal tiene la cualidad de que el valor R

cuadrado de su modelo nunca disminuirá con variables adicionales, solo iguales o superiores. Por qué se calcula con la cantidad de variables independientes, su modelo podría parecer más preciso con múltiples variables incluso si contribuyen de manera deficiente. El R-cuadrado ajustado penaliza la fórmula de R-cuadrado en función del número de variables, por lo tanto, una puntuación ajustada más baja puede indicarle que algunas variables no contribuyen correctamente al R-cuadrado de su modelo.

F-statistic prueba la hipótesis nula de que los coeficientes de la regresión sean igual a cero y la regresión no tiene posibilidades de predecir. La hipótesis nula se acepta cuando **prob (F-statistic)** es menor a 0.05, por lo tanto la hipótesis nula se rechaza.

Log-Likelihood es la región de rechazo de la función de máxima verisimilitud.

Akaike y Bayes son coeficientes que básicamente nos ayudaran a comparar contra otros modelos cuando las demás pruebas son similares. Mientras más bajos mejor.

Los valores de probabilidad de ambos coeficientes son también menores que 0.05 y por lo tanto aceptables.

Los intervalos de confianza indican con la confianza de 95% para ambos lados de la distribución de que los valores de los coeficientes sean $328.56 < \beta_0 < 365.499$ y $29.006 < \beta_1 < 31.107$

Omnibus y Jarque-Bera: prueban la normalidad basada en la simetría y curtosis de los errores. La hipótesis nula prueba la normalidad y se acepta cuando p-value es mayor que 0.05, se rechaza cuando los valores son menores a 0.05. Los dos indican que no existe normalidad porque sus probabilidades son valores menores de 0.05.

skew (sesgo) 0.82 indica que la distribución de los errores esta sesgado a la izquierda. 0 indica la distribución esta perfectamente centrada y simétrica.

kurtosis (curtosis) 4.2 es mayor a 3 por lo tanto ser una distribución leptocúrtica con valores concentrados en la media.

Durbin-Watson calcula valores entre 0 y 4 e indica independencia cuando son cercanos 2 o entre 1.5 y 2.5. Indica que no hay autocorrelación.

g) ¿Cuál sería el valor de T_tablas con el que contrastaría el valor de T_calculada si se tuviera un nivel de significancia del 0.05?
Los valores $t_0 = 36.8$ y $t_1 = 56.1$ tienen que compararse con 1.96 por que es el valor cuando $t_{n-1;\alpha/2}$

h) Indique si el modelo lineal se ajusta a los datos basado en las predicciones obtenidas. Justifique su respuesta.
Podría decirse que el modelo se ajusta un 31% pero no se cumple los supuestos de normalidad y homocedasticidad de los errores.

i) Si no existiera normalidad y homocedasticidad, ¿qué puede concluir de los resultados del análisis? ¿Qué solución propone ante la falta de normalidad y homocedasticidad?
La variable temperatura no es suficiente para consistentemente predecir la cantidad de bicis rentadas y el

modelo no es adecuado. Podemos (1) intentar con otra variable, (2) tratar con una regresión múltiple o, (3) efectuar una transformación de los datos de manera que los datos ya cumplan todas la hipótesis del modelo, como por ejemplo: La hora es importante para determinar la cantidad de bicicletas, usualmente no hay bicicletas rentadas en la madrugada o un *buen clima* es determinante.

j) ¿Qué son los "outliers"? ¿Cómo influyen en el análisis de regresión? ¿qué solución propone ante la presencia de estos valores?

Entiéndase outliers como los datos atípicos que ocurren con muy poca frecuencia, improbables, extraña, lejos del valor esperado. Sí buscamos normalidad los outliers están en los extremos de ambas colas. Afectan la regresión por que pueden generar un efecto desproporcionado en los resultados estadísticos y modifica los coeficientes de la ecuación de la regresión que puede conducir a interpretaciones engañosas. Pueden darse incluso por errores en la captura de datos, el proceso, la probabilidad de ocurrencia. Dependiendo de la naturaleza del atípico y la intención del análisis podrían ser corregidos o eliminados; o estudiados si la intención del análisis son los outliers, la baja probabilidad de ocurrencia, y los efectos que pueden tener.

k) Para el modelo de regresión lineal múltiple utilice la misma variable dependiente del inciso b). ¿cuáles son las variables independientes? (según la base de datos que eligieron)

- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Holiday - Holiday/No holiday

l) Escriba un enunciado planteando el problema a resolver.

Usando las variables independientes descritas arriba, encuentra un modelo de regresión lineal múltiple que prediga las bicicletas requeridas.

In [93]:
```
y = df[['Rented Bike Count']]
x = df[['Hour', 'Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)', 'Visibility (10m)',
        'Dew point temperature(°C)', 'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cr
X_train,X_test,Y_train,Y_test = train_test_split(x, y, test_size=0.2)
X = sm.add_constant(X_train, prepend=True)
rlm = sm.OLS(endog=Y_train, exog=X)
rlm = rlm.fit()
print(rlm.summary())
Y_pred = rlm.predict(X)
error = Y_train - Y_pred
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Rented Bike Count   R-squared:                       0.508
```

```
Model:                            OLS   Adj. R-squared:                  0.508
Method:                 Least Squares   F-statistic:                     699.4
Date:                Sun, 21 Nov 2021   Prob (F-statistic):               0.00
Time:                        18:19:52   Log-Likelihood:                -50990.
No. Observations:                6772   AIC:                         1.020e+05
Df Residuals:                    6761   BIC:                         1.021e+05
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          444.5194    105.953      4.195      0.000     236.818     652.221
Hour                            28.1347      0.858     32.783      0.000      26.452      29.817
Temperature(°C)                 29.9234      4.167      7.181      0.000      21.754      38.093
Humidity(%)                     -7.5550      1.177     -6.417      0.000      -9.863      -5.247
Wind speed (m/s)                 7.8187      5.995      1.304      0.192      -3.934      19.572
Visibility (10m)                 0.0382      0.011      3.369      0.001       0.016       0.060
Dew point temperature(°C)        2.1028      4.387      0.479      0.632      -6.497      10.702
Solar Radiation (MJ/m2)        -72.8732      8.936     -8.155      0.000     -90.391     -55.355
Rainfall(mm)                   -65.0268      5.234    -12.423      0.000     -75.288     -54.766
Snowfall (cm)                   15.5444     12.603      1.233      0.217      -9.161      40.250
Holiday                       -141.4544     25.524     -5.542      0.000    -191.489     -91.420
==============================================================================
Omnibus:                     1034.690   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2032.484
Skew:                           0.943   Prob(JB):                         0.00
Kurtosis:                       4.910   Cond. No.                     3.02e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.02e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R-squared un valor 0.51 en el coeficiente de determinación; mejor que el de la regresión lineal simple.

Prob (F-statistic) es menor a 0.05, por lo tanto la hipótesis nula se rechaza y los coeficientes son diferentes a 0

Akaike y Bayes son más bajos para estos modelos

Los valores de probabilidad Las variables *snowfall, wind speed y Dew point temperature* tienen un p-value mayor a 0.05 e indica que no son significativos. Todos los demás los coeficientes son menores que 0.05 y por lo tanto aceptables.

Omnibus y Jarque-Bera: ambos dos indican que no existe normalidad porque sus probabilidades son valores menores de 0.05.

skew (sesgo) 0.9 indica que la distribución de los errores esta sesgado a la izquierda.

kurtosis (curtosis) 4.8 es mayor a 3 por lo tanto ser una distribución leptocúrtica con valores concentrados cerca de la media.

Durbin-Watson calcula valores entre 0 y 4 e indica independencia cuando son cercanos 2 o entre 1.5 y 2.5. Indica que no hay autocorrelación.

o) ¿Qué modelo de regresión es mejor, el simple o el múltiple? Justifique su respuesta
En este caso la regresión múltiple es mejor, especialmente por la r ajustada, valores de akaike y bayes porque las pruebas de normalidad y homocedasticidad son prácticamente iguales.

Análisis de varianza

Voy a utilizar la base de datos SeoulBikeData.csv

a) ¿Qué supuestos se deben cumplir para realizar el ANOVA?

- La variable dependiente debe medirse al menos a nivel intervalo, o sea una variable de unidad de medida constante, los valores representan magnitudes, y la diferencia entre valores tienen un tamaño constante.
- Normalidad: Los valores de la muestra provienen de una distribución normal.
- Aleatoriedad: Los datos se seleccionan al azar.
- Independencia: Cada observación es independiente de cualquier otra observación.
- Homocedasticidad: Varianza constantes.

b) ¿Cuál es el objetivo del ANOVA?
Permite comparar las medias de varias poblaciones a partir del estudio de sus varianzas. Constituye la herramienta básica para el estudio del efecto de uno o más factores sobre la media de una variable continua

c) Indique las variables que va a utilizar y especifique quién es la variable dependiente y las variables independientes La variante independiente es la cantidad de bicicletas rentadas y las variables dependientes las voy a definir como:

- la temperatura: frío [-20 - 0], templado [0 - 20], caliente [20 - 40]
- es de día basado en la radiación solar: es de día cuando es mayor a 0, es noche cuando = 0.

d) Escriba un enunciado planteando el problema a resolver. (Plantear la H0 y H1 del problema a resolver).
Dada la sensación de temperatura y si es de día analiza la varianza de estos factores cuando se rentan bicicletas.
H_0 Las medias poblacionales del primer factor son iguales.
H_0 Las medias poblacionales del segundo factor son iguales.
H_0 No hay interacción entre los dos factores

Creo que la temperatura y la hora del día o en su defecto haya luz del sol afecta la cantidad de bicicletas rentadas.

e) ¿Cuántos niveles tiene cada factor que eligió?
Sensación de temperatura tiene tres niveles y si es de día o no tiene dos.

f) Obtenga el modelo del ANOVA y de una interpretación de los resultados.
El valor p obtenido del análisis de varianza es significante ($p < 0.05$) y por lo tanto concluyo que hay

diferencias significantes entre los factores. De cada uno de los factores y la iteración entre ambos, sensación de temperatura y si es de día.

g)¿Cuál sería el valor de F_tablas con el que contrastaría el valor de F_calculada si se tuviera un nivel de significancia del 0.05?

- El valor F en la tabla es 2.0838 cuando grados de libertad del numerador = 2 y los grados de libertad del denominador son 8462.
- El valor F en la tabla es 2.7055 cuando grados de libertad del numerador = 1 y los grados de libertad del denominador son 8463.

h) ¿Para que sirven las pruebas Post-Hoc en el ANOVA?
Después de determinar que las medias no son las mismas, las pruebas post hoc y las comparaciones múltiples por parejas permiten determinar qué medias difieren. Las pruebas de rango identifican subconjuntos homogéneos de medias que no se diferencian entre sí.

i) De ser necesario, incluya la prueba de Tukey y de la interpretación de los resultados.
Básicamente los resultados están indicando los intervalos de confianza para las diferencias entre las medias de los niveles de los factores. En todos los grupos se rechaza con un 95% de confianza ya que el p value es menor que 0.05.

j) Conclusiones generales de análisis de la varianza que acaba de realizar.
Los resultados obtenidos indican que los grupos hechos y seleccionados tienen varianza y las medias no son iguales y, la temperatura o si es de día o de noche es determinante en la renta.

In [123…
```
f = lambda x: 'caliente' if x >= 20 else 'frio' if x <= 0 else 'templado'
df['Sensacion_temperatura'] = df['Temperature(°C)'].map(f)
f = lambda x: '1' if x > 0 else '0'
df['dia'] = df['Solar Radiation (MJ/m2)'].map(f)
df['RentedBikeCount'] = df['Rented Bike Count']
anova_2 = ols('RentedBikeCount ~ C(Sensacion_temperatura) + C(dia) + Sensacion_temperatura:di
tabla_anova_2 = sm.stats.anova_lm(anova_2, typ=2)
print('2 Way ANOVA \n', tabla_anova_2)
```

```
2 Way ANOVA
                               sum_sq      df            F        PR(>F)
C(Sensacion_temperatura)   6.871144e+08     2.0  1236.139337  0.000000e+00
C(dia)                     1.075639e+08     1.0   387.021216  2.766359e-84
Sensacion_temperatura:dia  1.141384e+09     5.0   821.353783  0.000000e+00
Residual                   2.350989e+09  8459.0          NaN           NaN
```

In [132…
```
interaction_groups =  'Sensacion_temperatura' + df.Sensacion_temperatura.astype(str) + '&' +
comp = mc.MultiComparison(df["RentedBikeCount"], interaction_groups)
post_hoc_res = comp.tukeyhsd()
print('Tukey HSD - Multicomparison /n' , post_hoc_res.summary())
```

```
Tukey HSD - Multicomparison /n                                  Multiple Comparison of Means -
Tukey HSD, FWER=0.05
================================================================================================
======================
              group1                              group2                meandiff  p-adj    low
er      upper    reject
```

```
-----------------------------------------------------------------------------------------------
----------------------
Sensacion_temperaturacaliente&dia0 Sensacion_temperaturacaliente&dia1    408.078  0.001   34
8.9532   467.2029   True
Sensacion_temperaturacaliente&dia0     Sensacion_temperaturafrio&dia0  -706.7314  0.001  -77
4.7914  -638.6715   True
Sensacion_temperaturacaliente&dia0     Sensacion_temperaturafrio&dia1  -609.7105  0.001  -69
2.9073  -526.5137   True
Sensacion_temperaturacaliente&dia0 Sensacion_temperaturatemplado&dia0   -386.009  0.001  -44
3.6335  -328.3845   True
Sensacion_temperaturacaliente&dia0 Sensacion_temperaturatemplado&dia1   -82.6371  0.001  -14
1.4165   -23.8577   True
Sensacion_temperaturacaliente&dia1     Sensacion_temperaturafrio&dia0 -1114.8095  0.001 -117
4.2949 -1055.3241   True
Sensacion_temperaturacaliente&dia1     Sensacion_temperaturafrio&dia1 -1017.7886  0.001 -109
4.1303  -941.4468   True
Sensacion_temperaturacaliente&dia1 Sensacion_temperaturatemplado&dia0   -794.087  0.001  -84
1.2775  -746.8965   True
Sensacion_temperaturacaliente&dia1 Sensacion_temperaturatemplado&dia1  -490.7152  0.001  -53
9.3091  -442.1212   True
    Sensacion_temperaturafrio&dia0     Sensacion_temperaturafrio&dia1    97.0209 0.0119    1
3.5675   180.4743   True
    Sensacion_temperaturafrio&dia0 Sensacion_temperaturatemplado&dia0   320.7225  0.001   26
2.7281   378.7168   True
    Sensacion_temperaturafrio&dia0 Sensacion_temperaturatemplado&dia1   624.0943  0.001   56
4.9524   683.2363   True
    Sensacion_temperaturafrio&dia1 Sensacion_temperaturatemplado&dia0   223.7015  0.001   14
8.5158   298.8873   True
    Sensacion_temperaturafrio&dia1 Sensacion_temperaturatemplado&dia1   527.0734  0.001   45
0.9989   603.1479   True
Sensacion_temperaturatemplado&dia0 Sensacion_temperaturatemplado&dia1   303.3719  0.001    25
6.615   350.1288   True
-----------------------------------------------------------------------------------------------
----------------------
```

Análisis de Componentes Principales

Voy a utilizar la base de datos SeoulBikeData.csv

a) ¿Qué supuestos se deben de cumplir para realizar un Análisis de Componentes Principales?

- Variables correlacionadas o linealidad.
- Normalidad de cada una de las variables. Es esperado que se escalen las variables para preferiblemente tener media cero y dependiendo del caso desviación estándar uno.

b) ¿Cuál es el objetivo de realizar el análisis de componentes principales?
Cuando se enfrenta a un gran conjunto de variables, los componentes principales nos permiten resumir este conjunto con un número menor de variables representativas que explican colectivamente la mayor parte de la variabilidad en el conjunto original. El problema es tener muchas variables (gran dimensionalidad), la solución es hacer transformaciones lineales (baja dimensionalidad).
PCA es un enfoque no supervisado, ya que involucra solo un conjunto de características $X_1, X_2, \ldots, X_p$, y sin respuesta asociada. Además de producir variables derivadas para su uso en los problemas supervisados, PCA también sirve como una herramienta para la visualización de datos. También se puede utilizar como herramienta para completar los valores faltantes en un conjunto.

c) ¿Qué es lo que sucede si las variables que se están utilizando para realizar el PCA no están correlacionadas? Justifique su respuesta.
Podría depender de que tan baja correlación. Sí la correlación es 0 la matriz de correlación es una matriz identidad, la cual tiene eigenvalues de 1 para todas las columnas; el primer componente principal no representa más varianza que los siguientes componentes principales y terminarías quedándote con todas las variables. Por otro lado, puede ser que no sean importantes para el estudio y podrían ser eliminadas, pero no necesariamente ya que su información estaría representada en los componentes.

d) ¿Bajo qué circunstancias se recomienda hacer una estandarización de los datos?
Bajo la presencia de outliers.

e) ¿Qué representan los vectores propios de la matriz de varianzas y covarianzas en el análisis de componentes principales?
La matriz de covarianza de los datos observados está directamente relacionada con una transformación lineal de datos no correlacionados. Esta transformación lineal está completamente definida por los vectores y valores propios de los datos. La matriz define tanto la difusión (varianza) como la orientación (covarianza) de nuestros datos, entonces, si quisiéramos representar la matriz de covarianza con un vector y su magnitud, deberíamos encontrar el vector que apunte en la dirección de la mayor dispersión de los datos.

f) ¿Qué representan los valores propios de la matriz de varianzas y covarianzas en el análisis de componentes principales?
Los eigenvalores corresponden al cuadrado del factor de escala en cada dimensión. El vector propio más grande de la matriz de covarianza siempre apunta en la dirección de la varianza más grande de los datos, y la magnitud de este vector es igual al valor propio correspondiente

In [153…

```
df = pd.read_csv("C:/Users/nuno/OneDrive - ITESO/Ciencia de Datos/"
                 "analisis_estadistico_multivariado/code/SeoulBikeData.csv",
                encoding = 'unicode_escape')
to_drop = df.index[df["Functioning Day"] == "No"].tolist()
df.drop(to_drop, axis=0, inplace = True)
df.drop(['Functioning Day'], axis=1, inplace=True)
df['Holiday'].replace('No Holiday', 0, inplace=True)
df['Holiday'].replace('Holiday', 1, inplace=True)

df = df[['Temperature(°C)', 'Humidity(%)', 'Wind speed (m/s)', 'Visibility (10m)',
        'Dew point temperature(°C)', 'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Snowfall (cr

scaler = StandardScaler()
scaler.fit(df)
scaled_data = scaler.transform(df)
pca = PCA()
pca.fit(scaled_data)
pca_score = pd.DataFrame(data=pca.components_, columns=df.columns)
pca_score
```

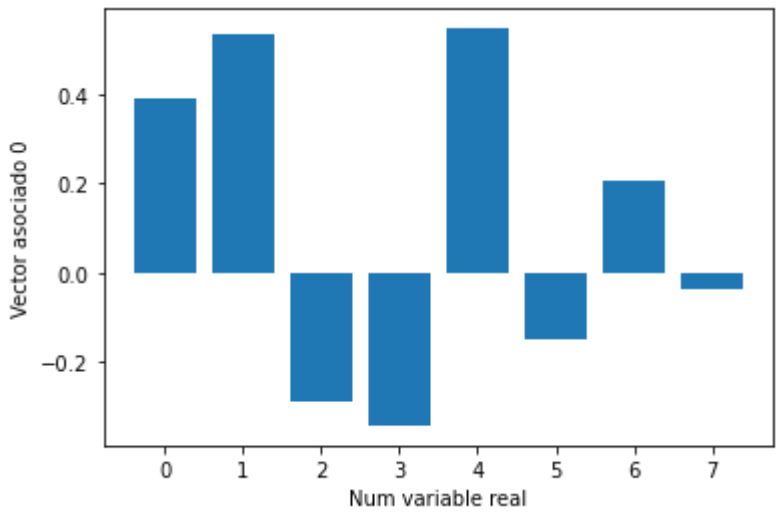
Out[153…

	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)

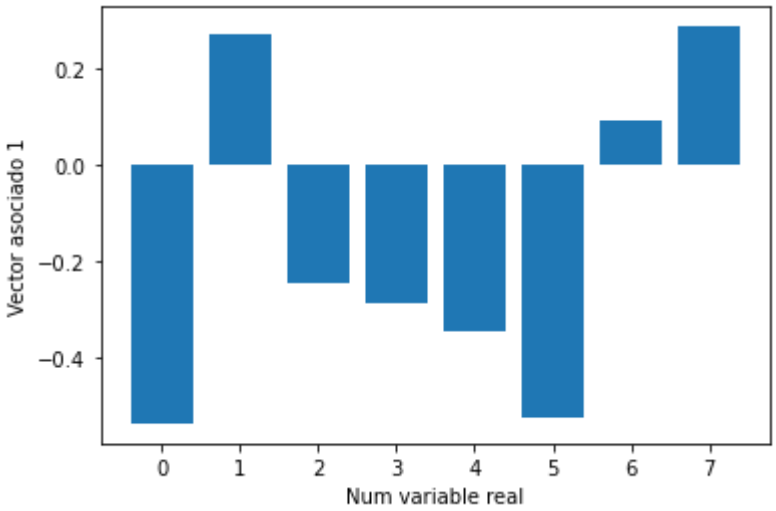
	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
0	0.391970	0.532504	-0.286126	-0.341907	0.546584	-0.147034	0.205594	-0.037272
1	-0.539073	0.272322	-0.247920	-0.290457	-0.347360	-0.526360	0.093604	0.287986
2	-0.015461	0.043659	0.523587	-0.327645	-0.021085	0.347854	0.495548	0.499345
3	-0.112563	-0.049772	0.089874	0.103169	-0.115762	-0.154594	0.721535	-0.638919
4	0.140491	-0.106071	-0.350292	0.652600	0.083887	-0.107660	0.394588	0.496077
5	0.055524	0.276247	0.671459	0.345446	0.174795	-0.541754	-0.157135	0.040891
6	0.357706	-0.685315	0.029586	-0.368051	0.050450	-0.500851	0.048247	0.101629
7	-0.626160	-0.284038	0.003090	-0.007629	0.725677	0.021998	0.008332	0.004250

In [154…

```
matrix_transform = pca.components_.T
plt.bar(range(df.shape[1]),matrix_transform[:,0])
plt.xlabel('Num variable real')
plt.ylabel('Vector asociado 0')
plt.show()
loading_scores = pd.DataFrame(pca.components_[0], index=df.columns)
sorted_loading_scores = loading_scores[0].abs().sort_values(ascending=False)
top_variables = sorted_loading_scores[0:3].index.values
print(top_variables)

plt.bar(range(df.shape[1]),matrix_transform[:,1])
plt.xlabel('Num variable real')
plt.ylabel('Vector asociado 1')
plt.show()
loading_scores = pd.DataFrame(pca.components_[1], index=df.columns)
sorted_loading_scores = loading_scores[0].abs().sort_values(ascending=False)
top_variables = sorted_loading_scores[0:3].index.values
print(top_variables)

per_var=np.round(pca.explained_variance_ratio_*100, decimals=1)
porcent_acum = np.cumsum(per_var)
porcent_acum
```



```
['Dew point temperature(°C)' 'Humidity(%)' 'Temperature(°C)']
```




```
['Temperature(°C)' 'Solar Radiation (MJ/m2)' 'Dew point temperature(°C)']
```

Out[154...
```
array([ 30.2,  54.7,  67.9,  79.6,  89. ,  97.1,  99.9, 100. ])
```

g) Obtenga la matriz de vectores propios y de una interpretación de los resultados.

En valores absolutos, mientras más grande el valor, mayor influencia tiene en el componente principal observado.

h) En el análisis de componentes principales que está realizando ¿Cuál es número óptimo de componentes principales? Justifique su respuesta

Usando una propuesta de Pareto 80-20, el número optimo son los 4 componentes principales.

i) ¿Cuáles son las variables que más influyen en los primeros dos componentes principales?

- primer componente principal Humidity(%), Dew point temperature(°C), Visibility (10m).
- segundo componenete principal Temperature(°C), Solar Radiation (MJ/m2), Dew point temperature(°C).

j) Conclusiones generales del análisis de componentes principales que acaba de realizar.

PCA en práctica es muy simple, pero lleva una fuerte carga de entendimiento a estadística, geometría euclidiana, algebra lineal. El análisis se enfocó primordialmente en las variables climatológicas y pues básicamente se crearon variables en ese sentido.

Como PCA es una herramienta de exploración no-supervisada y reducción de dimensionalidad, obtuvimos valores que podrían ser utilizados para regresión de componentes principales para ser usado para predecir la cantidad de bieletas.